information

MDPI

Article

A Discriminative Framework for Action Recognition Using f-HOL Features

Samy Bakheet [1,*] and Ayoub Al-Hamadi [2]

[1] Department of Math and Computer Science, Faculty of Science, Sohag University, 82524 Sohag, Egypt

[2] Institute for Information Technology and Communications, Otto-von-Guericke-University Magdeburg, P.O. Box 4120, 39016 Magdeburg, Germany; ayoub.al-hamadi@ovgu.de

[*] Correspondence: samy.bakheet@gmail.com or sbakheet@ovgu.de; Tel.: +49-391-67-11481

Academic Editor: Willy Susilo

Received: 30 August 2016; Accepted: 8 November 2016; Published: 22 November 2016

Abstract: Inspired by the overwhelming success of Histogram of Oriented Gradients (HOG) features in many vision tasks, in this paper, we present an innovative compact feature descriptor called fuzzy Histogram of Oriented Lines (f-HOL) for action recognition, which is a distinct variant of the HOG feature descriptor. The intuitive idea of these features is based on the observation that the slide area of the human body skeleton can be viewed as a spatiotemporal 3D surface, when observing a certain action being performed in a video. The f-HOL descriptor possesses an immense competitive advantage, not only of being quite robust to small geometric transformations where the small translation and rotations make no large fluctuations in histogram values, but also of not being very sensitive under varying illumination conditions. The extracted features are then fed into a discriminative conditional model based on Latent-Dynamic Conditional random fields (LDCRFs) to learn to recognize actions from video frames. When tested on the benchmark Weizmann dataset, the proposed framework substantially supersedes most existing state-of-the-art approaches, achieving an overall recognition rate of 98.2%. Furthermore, due to its low computational demands, the framework is properly amenable for integration into real-time applications.

Keywords: human action recognition; Histogram of Oriented Gradients (HOG); fuzzy Histogram of Oriented Lines (f-HOL); Latent-Dynamic Conditional random fields (LDCRFs); Weizmann action dataset

1. Introduction

In recent years, automatic recognition of human activities from both still images and video sequences has attracted tremendous research interest, due to its immense potential for many applications in various fields and domains [1,2]. Although many efficient applications are available for the purpose of human action recognition, the most active widespread application might be Human Computer Interaction (HCI), where no explicit user's actions (e.g., keystrokes and mouse clicks) are available to capture user input. Instead, interactions are more likely to occur through human actions and/or gestures [3]. In this sense, it is worth pointing out an indisputable fact that the recognition of human actions is an effortless process for us as human beings, but a very challenging task for computers. The task of action recognition shares many challenges with other tasks in computer vision and pattern recognition, such as object detection and tracking, motion recognition, etc. The major common challenges in human action recognition shared with other problems in computer vision include illumination conditions, occlusions, clutter in background, object deformations, intra/inter class variations, pose variations, and camera point-of-view.

Furthermore, while human vision has an extraordinary ability to efficiently recognize human actions from video data, with a high degree of accuracy, it is an arduous task for the computer to do such a task in a very similar manner. First, the fact that the same action is performed by different people at different velocity poses a quite serious technical problem to any automatic striving to achieve action recognition task optimally. Moreover, moving shadows generated by bad lighting conditions can also degrade tracking the motion of human body parts. Some body parts can be occluded owing to camera viewpoints that provides an additional difficulty to the human action recognition task. Added to that, small moving objects (i.e., distractors) in background are also another problematic issue for this task. For example, in a scene of crowded street, trees swinging and/or shop advertisements blinking in the background are challenging issues for motion detection and tracking.

It is worth emphasizing here that, while the early research into human motion modeling and/or recognition dates back to the pioneering work of Johansson [4] in the mid-1970s, research on human action recognition did not shift to the forefront until the early 1990s. About a decade later, by the end of the 1990s, research in human action recognition has barely begun to get into its infancy [5,6]. It may be interesting to state that the past ten years or so have witnessed an increasing number of research efforts in human action recognition. However, the contributions to improved human action recognition have been modest, as well as the off-the-shelf technology solution space for human action recognition is still far from being quite mature yet. The experimental systems of action recognition are now appearing at a very limited number of locations (e.g., airports and other public places).

The primary objective of this paper is to perform an innovation design for improving the recognition of human actions in video sequences, by developing an efficient and reliable method for modeling and recognizing human actions from video sequences. In our approach, we present a new feature descriptor called fuzzy Histogram of Oriented Lines (f-HOL) for action recognition, which is a distinct variant of Histogram of Oriented Gradients (HOG) features. The extracted features are then used by a discriminative Latent-Dynamic Conditional random fields (LDCRFs) model to recognize actions from the video frames. A set of validation experiments are conducted on the benchmark Weizmann action recognition dataset. The preliminary results achieved with this approach are promising and compare very favorably to those of other investigators published in the literature.

The rest of this paper proceeds as follows. In Section 2, we present an overview of related work to provide relevant background knowledge concerning the problem domain. The architecture of the proposed framework for human action recognition is fully detailed in Section 3. Section 4 summarizes our extensive experiments comparing the results achieved by the proposed approach with those of other similar state-of-the-art methods from literature. Finally, we conclude and suggest possible directions for future research in Section 5.

2. Related Literature

Over the course of nearly two decades or so, a large amount of literature has been reported by many researchers in the fields of computer vision and pattern recognition for video-based human action recognition [3,7–11], motivated by a wide spectrum of real-world applications, such as intelligent human–computer interface, detection of abnormal events, video retrieval, autonomous video surveillance, etc. Broadly speaking, human actions can be recognized using various visual cues, such as motion [12,13] and shape [14]. Extensive literature surveys reveal that there exists an increasing corpus of prior work on human action recognition focusing on using spatial-temporal keypoints and local feature descriptors [1,15,16]. The local features are extracted from the region around each keypoint detected by the keypoint detection process. These features are then quantized to provide a discrete set of visual words before they are fed into the classification module.

In [9], Blank et al. define actions as 3D space-time shapes generated by accumulating the detected foreground human figures, while they assume fixed camera and known background appearance. Their method greatly depends on moving object tracking, where various space-time features (e.g., local space-time saliency, action dynamics, shape structures, and orientation) are extracted for action recognition. In addition, there is another thread of research targeted at analyzing patterns of motion to recognize human actions. For instance, in [17], the authors analyze the periodic structure of

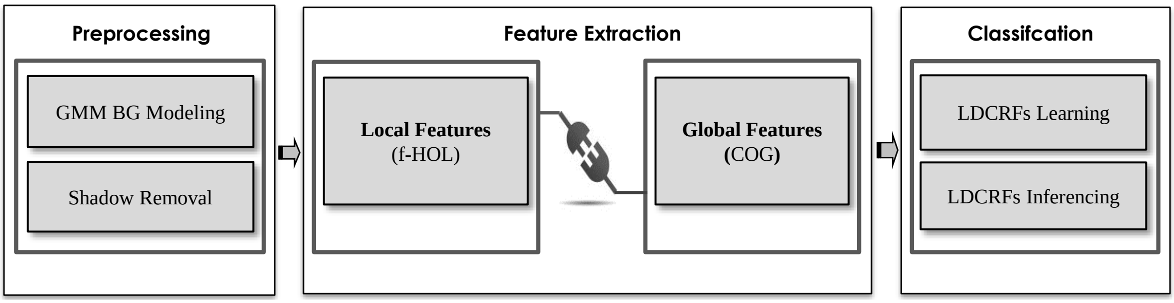

Figure 1. A schematic diagram of the proposed methodology for action recognition.

optical flow patterns for gait recognition. In the same vein, in [18], periodic motions are detected and classified to recognize actions from video sequences.

Alternatively, several researchers have proposed using both motion and shape cues. For example, in [19], the authors detect the similarity between video segments using a space-time correlation model. In [20], Bobick and Davis use temporal templates, including motion-energy images and motion-history images to recognize human movement. Rodriguez et al. [21] present a template-based approach using a Maximum Average Correlation Height (MACH) filter to capture intra-class variabilities. Moreover, in a more recent work by Fernando et al. [22], an effective approach for action recognition is proposed, which has the potential to adequately represent action videos using a ranking machine. More specifically, given the frame descriptors, the action video can then be represented by the hyperplane that ranks the frames based on their temporal orders.

Furthermore, in [23], Kantorov and Laptev exploit optical flow measurements from videos to encode the pixel motions. In their work, as a spatiotemporal descriptor for action representation, they present the Histogram of optical Flow (HoF) over local regions. Since the measurement of optical flow is computationally intensive, the authors opted for the use of video decompression techniques. More specifically, they opted for not obtaining the HoF descriptor (or its more recent extension MBH descriptor) from the estimation of original optical flow fields. Instead, they make use of the motion fields in MPEG compression. This motion field, the so-called MPEG Flow, can be generated virtually free during a video decoding operation.

Furthermore, a substantial amount of research has been conducted on the modelling and understanding human motions by constructing elaborated temporal dynamic models [24,25]. In addition, there is an increasing body of research reporting promising results using generative topic models for visual recognition based on the so-called Bag-of-Words (BoW) models. The key concept of a BoW is that the video sequences are represented by counting the number of occurrences of descriptor prototypes, so-called “visual words”. Topic models are built and then applied to the BoW representation. Three of the most popularly used topic models are Correlated Topic Models (CTM) [26], Latent Dirichlet Allocation (LDA) [27] and probabilistic Latent Semantic Analysis (pLSA) [28].

3. Proposed Methodology

In this section, the proposed approach for video-based action recognition is described. A step-by-step overview of the proposed methodology is presented in Figure 1. As schematically illustrated in the above figure, the general framework of our proposed approach proceeds as follows. The first step consists in separating moving objects (i.e., human body parts) from the background in a given video sequence. Adaptive background subtraction is most appropriate to achieve this goal of motion detection. The segmented body parts can further be refined by standard morphological operations, such as erosion and dilation. Then, a set of low dimensional local features is extracted from the silhouettes of moving objects. Finally, a 3D vector representation is formed from the extracted features and then fed into an LDCRF classifier for action classification. The details of each part of the proposed technique are described in the remainder of this section.

3.1. Background Subtraction and Shadow Removal

In this step, we aim at robust background subtraction involving shadow detection and removal to track human movements in video sequences. To achieve this goal, we employ an effective algorithm for background modeling, subtraction, update, and shadow removal [29]. The fundamental idea of the algorithm consists in modelling the background color distribution with adaptive Gaussian mixtures, coupled with color-based shadow detection. Adaptive Gaussian mixtures are employed in the combined input space of luminance-invariant color to distinguish moving foreground from their moving cast shadows in video sequences.

In its most general form, background subtraction is a commonly used paradigm for detecting moving objects in a scene taken from a stationary camera, which involves two distinct processes that operate in a closed loop: background modeling and foreground detection. A standard approach for background modeling involves the construction of a model for background in the field of view of a camera. Then, the background model is periodically updated to account for illumination changes. In the foreground detection, a decision is used to ensure that the model is fit to the background. The resulting change label field is fed back into background modeling so that no foreground intensities contaminate the background model. With respect to Gaussian mixtures, it is perhaps not irrelevant to point out that Gaussian Mixtures Models (GMMs) are an instance of a larger class of density models that have several functions as additive components [30].

In this work, Gaussian mixture models are used for modelling background. On this model, each pixel in the scene is modeled using a mixture of K (usually set from three to five) Gaussian distributions; we used K = 3 in our experiments. The persistence and variance of each Gaussian of the mixture are used to determine which Gaussian probably corresponds to background colors. Pixels whose color values do not fit the background distributions are detected as part of the foreground or moving objects. More formally, let $\{X_1, \dots, X_t\}$ be the history associated with a pixel at time t, where $X_i (i = 1, \dots, t)$ are measurements of the RGB (Red, Green, Blue) values at time i. The recent history of each pixel can be modeled reasonably well with a mixture of K Gaussian distributions. Thus, the probability of observing the current pixel value is defined as follows:

$$P(X_t) = \sum_{i=0}^{\mathrm{K}} \omega_{i,t}\, \eta(X_t, \mu_{i,t}, \Sigma_{i,t}), \tag{1}$$

where $\omega_{i,t}$, t. $\mu_{i,t}$ and $\Sigma_{i,t}$ are an estimate of the weight, the mean value, and the covariance matrix of the i-th Gaussian in the mixture at time t, respectively. η is a Gaussian probability density function:

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t - \mu_t)^T \Sigma^{-1} (X_t - \mu_t)}. \tag{2}$$

Assuming the independence of the color channels, $\Sigma_{i,t}$ can be expressed as: $\Sigma_{i,t} = \sigma_i^2 \cdot I$. Thereafter, online approximation is used to update the model in an iterative manner as follows. At each pixel, all parameters of the most matched Gaussian are updated via an online K-means approximation, whereas only the weight parameters of others are updated, while their means and variances remain unchanged (full details about model parameter's updates can be found in [5]). Figure 2 shows a sample of results for the human object scene when the video segmentation technique with Gaussian mixture model (GMM) and shadow removal is applied, where the moving object along with the shadow is extracted.

It is generally recognized that an effective design of a color model that is able to separate the brightness from the chromaticity component can robustly remove shadow from images [31]. Formally speaking, for a given pixel, let $B_t = [\mu_r, \mu_g, \mu_b]$ be the expected background value approximated from

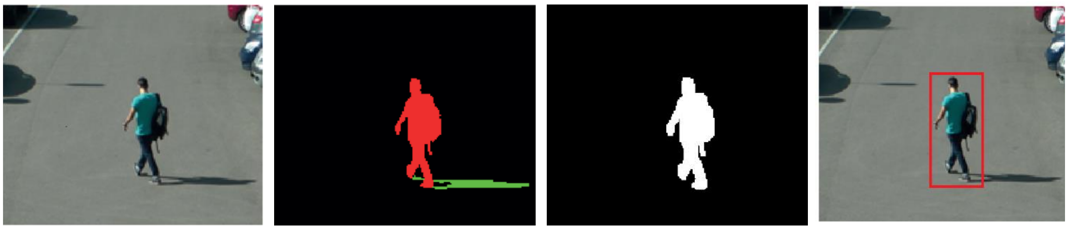
(a) (b) (c) (d)

Figure 2. An example of shadow and motion detection in a real outdoor video: (**a**) an input frame; (**b**) subtracted background and detected shadow, where foreground pixels are marked in **red** color, whereas shadow pixels are marked in **green** color; (**c**) silhouette after removing shadow; and (**d**) moving object detection.

a set of first training frames. Thus, in each coming frame $X_t = [x_r(t), x_g(t), x_b(t)]$, brightness α_t and chromaticity distortions χ_t can be computed from the background value as follows:

$$\begin{aligned} \alpha_t &= \frac{\left(\frac{x_r(t)\mu_r}{\sigma_r^2} + \frac{x_g(t)\mu_g}{\sigma_g^2} + \frac{x_b(t)\mu_b}{\sigma_b^2}\right)}{(\frac{\mu_r}{\sigma_r})^2 + (\frac{\mu_g}{\sigma_g})^2 + (\frac{\mu_b}{\sigma_b})^2}, & (3) \\ \chi_t &= \sqrt{(\frac{x_r - \alpha_t\mu_r}{\sigma_r}\Big)^2 + \Big(\frac{x_g - \alpha_t\mu_r}{\sigma_g}\Big)^2 + \Big(\frac{x_b - \alpha_t\mu_r}{\sigma_b}\Big)^2}. & (4) \end{aligned}$$

From a geometrical point of view, in RGB space, we can see the chromaticity distortion χ_t as the length of the vector that goes from the pixel value X_t to the plane perpendicular to the line connecting the background value μ with the zero intensity point. It is a scalar that can be thought of as a measure of indicating how much the pixel color varies from the background color. The chromaticity distortion can be simply normalized with division by its variation evaluated during the training phase. Formally, the normalization process is defined as follows:

$$\hat{\chi}_t = \frac{\chi_t}{v},\; v = \sqrt{\frac{\sum_{t=1}^{\tau} \chi_t^2}{\tau}}, \tag{5}$$

where τ denotes the number of training frames. In this context, a pixel noise is used to normalize the variation between the pixel value and the background value. Hence, the full statistical model allows for the application of only one single threshold τ_χ for variation. Following the original approach gives a sense of a proper value for the threshold by plotting the histogram of the normalized chromaticity distortion during the training sequence, exempt of foreground perturbations. In practice, the successful detection rate experienced by the user plays a crucial role in determining the optimum value for τ_χ.

Depending upon their normalized brightness distortions and normalized chromaticity distortions, pixels are then categorized into four clusters: labeled background, cast shadow, highlight or foreground, as follows. Pixels having small normalized brightness distortion and small normalized chromaticity distortion belong to the background, whereas, pixels that have a small normalized chromaticity distortion and a lower or higher brightness value than the background value are labeled as cast shadow or highlight, respectively. All of the remaining unclassified pixels are automatically assigned to foreground objects or moving objects. More specifically, a pixel is detected as a cast shadow if it meets the following coupled constraints:

$$\hat{\chi}_t < \tau_\chi \wedge \alpha_{min} < \alpha_t < 1, \tag{6}$$

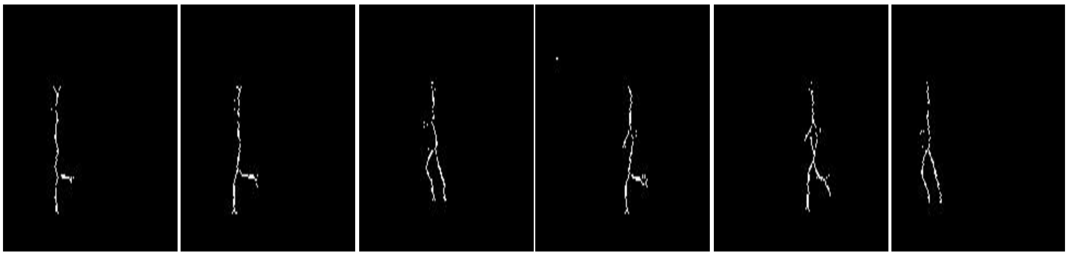

Figure 3. Sample sequence of skeleton images for a video of running action.

where ∧ denotes "and". After shadow elimination, there may remain some small isolated regions and noises in the binary image. The extracted foreground objects are usually affected by these noises and artifacts. Therefore, a median filter is first applied to suppress small artifacts and remove noises. Afterwards, an adaptive "close-opening" filter (i.e., made up of closing operator followed by opening operator) is applied to improve the quality of segmentation and preserve the edges of the moving objects. The structuring elements of size 3 × 3 and 5 × 5 were tried. Finally, a typical blob analysis process scans through the entire segmented image to detect all of the moving objects (or blobs) in the image and builds a detailed report on each moving object [32].

3.2. Feature Extraction

For successful action feature extraction, the accurate segmentation of the action silhouette points is essential. In this section, we present an innovative compact representation for action recognition based on skeleton dynamics that collects multiple action information such as static pose, motion, and overall dynamics. In order to represent a human action entirely, we propose mainly capturing the body shape variations based on action skeleton information. In achieving this goal, in our approach, we mainly focus on the gradients of one-dimensional line representation of action skeletons, namely dynamic features.

Generally, the skeletonization process aims at reducing foreground regions (i.e., objects of interest) in a segmented binary image to a skeletal remnant that maintains the connectivity of objects in the original image. The skeleton is "central-spine", which can easily be shown to be a 1D line representation of a 2D object. It is easy to demonstrate that the skeleton retains the topology of the original shape, and the original shape can be reconstructed from its own skeleton. In order to find the skeleton of a given silhouette image, a set of morphological operations is applied to the segmented silhouette image. These operations include multiple morphological dilation operations. Then, the same number of morphological erosion operations are applied. There exist some holes that were generated during the segmentation process. These holes are filled thanks to the application of the dilation operation followed by the erosion operation to improve the skeletonization process and restore the shape of the human body. As a result, we obtain the skeletons of moving human body parts. More specifically, we designed the following code snippet mainly based on the real-time OpenCV library and C++ interface to extract the skeleton from the segmented video sequence:

```
Mat element = getStructuringElement(MORPH_CROSS, Size(3, 3));
do
{
  erode(img, eroded, element);
  dilate(eroded, temp, element);
  subtract(img, temp, temp);
  bitwise_or(skel, temp, skel);
  eroded.copyTo(img);
  done = (norm(img) == 0);
} while (!done).
```

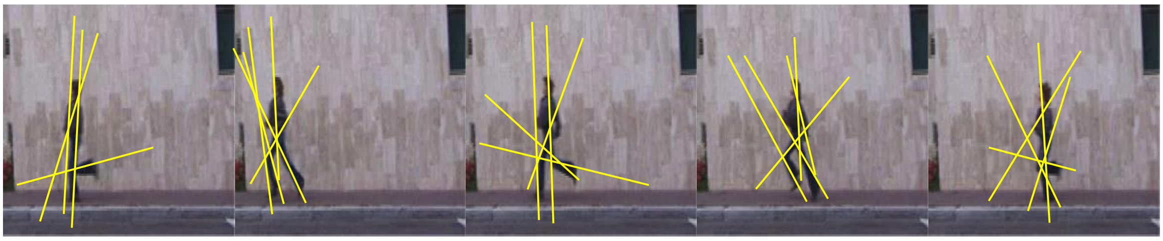

Figure 4. An example of straight-line detection in a video sequence of running action.

Figure 3 shows a sample sequence of skeleton images for a video sequence of a running action. Once binary skeletons of a given action video have been obtained, the probabilistic Hough transform algorithm described in [33] can be employed to detect straight-lines in each frame of the action video. An example of straight-line detection in a video of running action is given in Figure 4. It would be worthy of note at this point that our designed function for line detection is not as expensive as other traditional OpenCV Hough detectors.

For local feature extraction, each video clip (i.e., action snippet) is firstly divided into several time slices defined by linguistic intervals. Gaussian functions are used to describe these intervals:

$$\mu_j(t;\varepsilon_j) = e^{-\frac{1}{2}\left|\frac{t-\varepsilon_j}{\sigma}\right|^m}, j = 1,2,\ldots,s, \tag{7}$$

where ε_j, σ, and m are the center, width, and fuzzification factor of temporal slices, respectively, and s is the total number of the time slices. It is important to note that all Gaussian membership functions defined earlier are chosen to be of the same shape such that their final cumulative sum is always unity for every instant of time. In this work, we aim at introducing a new action representation based on computing rich descriptors from detected straight-lines that capture more local spatiotemporal information. Human action is generally composed of a sequence of temporal poses. Thus, reasonable estimate of an action pose can be constructed from a finite set of detected lines.

Formally, let L be a set of line segments detected using the probabilistic Hough transform from an action snippet at a time instant t. Assuming each line is represented by a four-element vector (x_1, y_1, x_2, y_2), where (x_1, y_1) and (x_2, y_2) are the ending points of each detected line segment, and then the magnitude and orientation of the gradient are calculated as:

$$\begin{aligned} \rho &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, \\ \theta &= \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right). \end{aligned} \tag{8}$$

In order to form a local feature descriptor, the so-called fuzzy Histogram of Oriented Lines (f-HOL) is constructed for a given action snippet. In order to achieve this, a separate f-HOL is computed for each time slice as follows:

$$h_j(k) = \sum_{\substack{\theta \in bin(k) \\ \rho > \rho_m}} \mu_j(t), j = 1,2,\ldots,s, \tag{9}$$

where ρ_m is a predetermined threshold to remove gradients of small magnitude. All the resulting 1D histograms are then normalized to achieve robustness to scale variations. Finally, these normalized histograms are concatenated into a single histogram to form the local feature vector of the action snippet. Figure 5 displays a sample of three action sequences from the Weizmann action dataset and their corresponding smoothed f-HOL descriptors. From top to bottom, the actions are “Bend”, “Jack” and “Jump”, respectively.

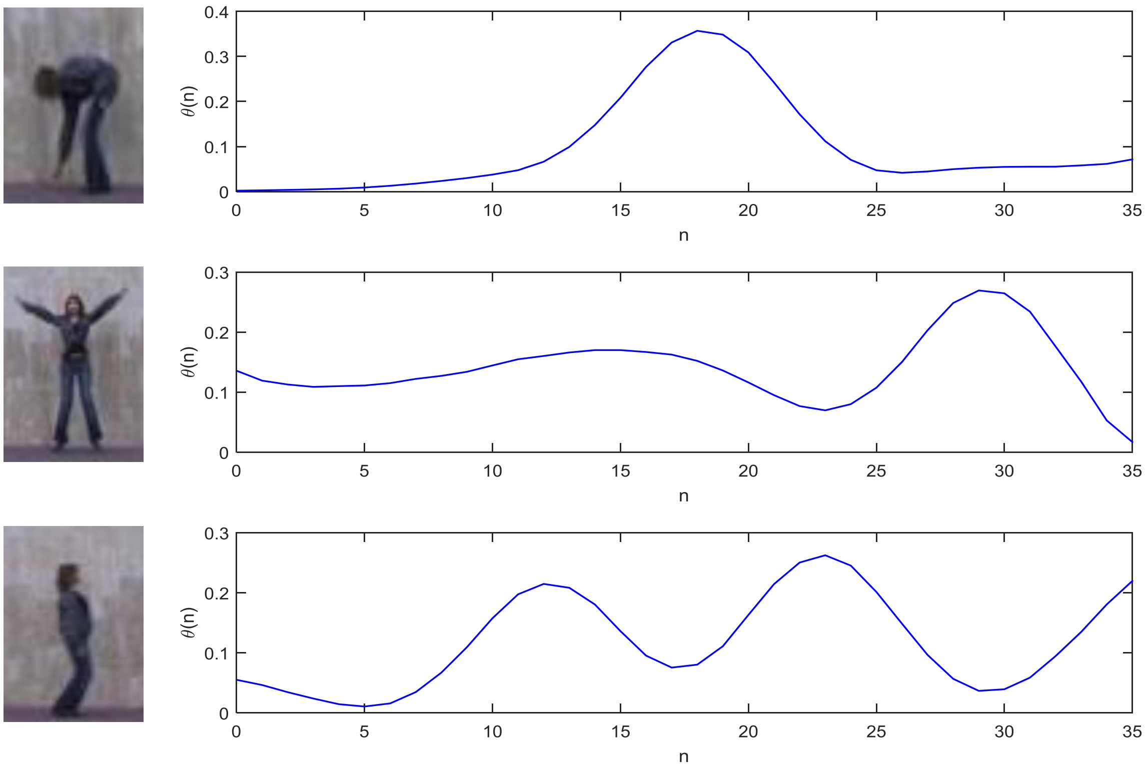

Figure 5. A sample of three actions along with their corresponding smoothed f-HOL features; from **top** to **bottom**, the actions are "Bend", "Jack", and "Jump", respectively.

Fusion of Local and Global Action Features

It follows from the discussion in the previous subsection that the local features extracted from action snippets using f-HOL have been highlighted. On the other hand, global features have been extensively applied to a wide range of object recognition problems and obtained surprisingly good results. This, in turn, fosters a strong motivation in us to extract global features and fuse them together with local features to build a more expressive and discriminative action representation. The extracted global features are based on computing the center of gravity (COG) that delivers the center of motion. Hence, we can obtain the motion sequence from the COG trajectory of the motion, where the center of motion is given by

$$\bar{z}(t) = \frac{1}{n}\sum_{i=1}^{n} z_i(t), \tag{10}$$

where $z_i(i = 1, 2, \ldots, n)$ are moving pixels in the current frame. It is of interest to note that the global features are potentially informative not only about the type of motion (e.g., translational or oscillatory), but also about the rate of motion (i.e., velocity). In addition, in our experiments, these features exhibit sufficient discriminative capabilities to distinguish, for example, between an action where motion occurs over a relatively large area (e.g., running) and an action localized in a smaller region, where only small body parts are in motion (e.g., boxing).

3.3. Action Classification

In this section, the details of the feature classification module in our action recognition system are described. Generally, the main purpose of the classification module in the current action recognition system is to classify a given action into one of a set of predefined classes, depending on the extracted features. The classification module depends on the availability of a set of previously labeled or classified actions. In this case, this set of actions is termed the training set and the resulting learning strategy

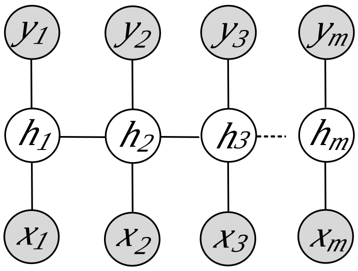

Figure 6. Graphical representation of the Latent-Dynamic Conditional Random Fields (LDCRFs) model. x_j represents the j-th observation, h_j denotes the hidden state assigned to x_j, and y_j is the class label of x_j. The nodes in **gray** circles represent the observed variables.

is called supervised learning. For the task of action classification, there are plenty of classification techniques reported in literature, including Naïve Bayesian (NB), k-Nearest Neighbor (k-NN), Support Vector Machines (SVMs), Neural Networks (NN), Conditional Random Fields (CRFs), etc. In this work, we opted to choose the Latent-Dynamic Conditional Random Fields (LDCRFs) for action classification. Due to their inherent dependence on CRFs, LDCRFs are characterized as discriminative models that have the capability to describe the substructure of a label and learn dynamics between labels. Moreover, it was found that LDCRFs perform well when applied to several large scale recognition problems, and they are superior to other learning methods (e.g., Hidden Markov Models (HMMs)) at learning relevant context and integrating it with visual observations [34,35].

Historically, LDCRF have been conceived as an improved extension to CRFs to learn the hidden interaction between features, and they can be interpreted as undirected graphical models capable of labeling sequential data. Therefore, they can be applied directly to sequential data avoiding the need for windowing the signal. In this manner, each label (or state) suggests a specific gesture. As LDCRFs include a class label for each observation, they are able to classify unsegmented gestures. Furthermore, the LDCRF model can perfectly infer the action sequences in the training and test phases.

Formally speaking, the basic task of the LDCRF model, as described by Morency et al. [36], is to learn a mapping between a sequence of observations $\mathbf{x} = \langle x_1, x_2, \ldots, x_m \rangle$ and a sequence of labels $\mathbf{y} = \langle y_1, y_2, \ldots, y_m \rangle$. Each y_j is a class label for the j-th observation in a sequence and is a member of a set $\mathcal{Y}$ of possible class labels. A feature vector $\phi(\mathbf{x}_j) \in \mathbb{R}^d$ is used to represent each image observation x_j. For each sequence, let $\mathbf{h} = \langle h_1, h_2, \ldots, h_m \rangle$ be a vector of substructure variables not observed in the training examples. Hence, these variables form a set of "hidden" variables in the model, as shown in Figure 6. Given the above definitions, a latent conditional model can thus be formulated as follows:

$$p(\mathbf{y}|\mathbf{x}, \theta) = \sum_{\mathbf{h}} p(\mathbf{y}|\mathbf{h}, \mathbf{x}, \theta) p(\mathbf{h}|\mathbf{x}, \theta), \tag{11}$$

where θ is a set of the model parameters. Given a set of sequences, each labeled with its correct class name $\{(\mathbf{x_i}, \mathbf{y_i}), i = 1 \ldots n\}$, the training objective is then to learn the model parameters θ using the following objective function [37]:

$$L(\theta) = \sum_{i=1}^{n} \log p(\mathbf{y}_i|\mathbf{x}_i, \theta) - \frac{1}{2\sigma^2} \|\theta\|^2, \tag{12}$$

where n is the total number of training sequences. It can be seen from the above equation that the first term on the right-hand side of the equation represents the conditional log-likelihood of the training samples, whereas the second term is the log of a Gaussian prior with variance σ^2, i.e.,

$$p(\theta) \sim \exp(\frac{1}{2\sigma^2} \|\theta\|^2). \tag{13}$$

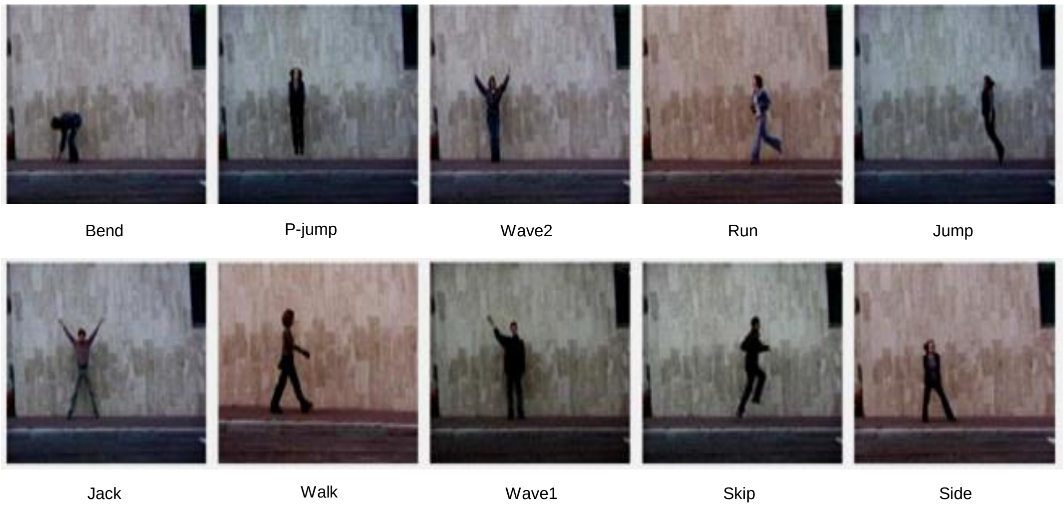

Figure 7. Sample actions from the benchmark Weizmann dataset.

In order to estimate the optimal model parameters, an iterative gradient ascent algorithm is employed to maximize the objective function:

$$\theta^* = \arg\max_{\theta} L(\theta). \tag{14}$$

Once the model parameters θ^* are learned, given an unseen (test) sample $\mathbf{x}$, the predicted class label y^* can be straightforwardly obtained via inference in the model as follows:

$$y^* = \arg\max_{y} p(y|\mathbf{x}, \theta^*). \tag{15}$$

For further details concerning the training and inference of LDCRF, the interested reader is referred to the full description given in [36].

4. Experiments and Results

In this section, our intensive experiments conducted to evaluate the proposed approach for action recognition are described and the obtained results are discussed. We start with a brief description of the action recognition dataset and the evaluation protocol that have been exploited to assess the performance of the recognition framework. After introducing the experiment settings and the evaluation protocol, we compare the proposed approach to related existing recognition methods. In the present work, the proposed recognition framework has been tested on the Weizmann dataset, which is regarded as one of the most widely used datasets for action recognition. This dataset was first presented by Blank et al. [9] in 2005, and, thereafter, it was made publicly available to researchers without a restriction or other access charge. The Weizmann dataset consists of a total of 10 action categories, namely "walk", "run", "jump-forward-on-two-legs" (or shortly "jump"), "jumping-in-place-on-two-legs" (or "p-jump"), "jumping-jack" (or "jack"), "gallop side ways" (or "side"), "bend", "skip", "wave-one-hand" (or "wave1") and "wave-two-hands" (or "wave2'). Each action is performed by nine persons. The action sequences were captured with a static camera over static background at a rate of 25 fps, with a relatively low spatial resolution of 180 × 144 pixels, 24 bits per pixel color depth. It should be noted that the action video clips (i.e., so-called action snippets) are of short duration; each snippet generally lasts only for a short period of time, namely just a few seconds. A sample frame for each action in the Weizmann dataset is shown in Figure 7.

In order to provide an unbiased estimate of the generalization abilities of the proposed method, the leave-one-out cross-validation (LOOCV) technique was applied for the validation process. As the name suggests, this involves using a set of action sequences from a single subject in the dataset as the testing data, and the remaining sequences as the training data. This is repeated such that each set of sequences in the dataset is used once as the validation. More specifically, the sequences of eight subjects were used for training and the sequences of the remaining subject were used for validation data. The LDCRF classifier is trained on the training set, while the evaluation of the recognition performance is performed on the test set. The recognition results achieved on the Weizmann dataset are fully summarized in form of a confusion matrix given in Table 1. From the figures in the matrix above, a number of interesting observations can be drawn. First, the majority of actions are correctly classified. In addition, an average recognition rate of 98.2% is achieved. Furthermore, there is a clear distinction between arm actions and leg actions. The vast majority of mistakes where confusions occur are only between skip and jump actions and between jump and run actions. This is also due to the high closeness or similarity among the actions in each pair of these action categories.

Table 1. Confusion matrix of Weizmann actions.

Action	Walk	Run	Jump	P-Jump	Jack	Side	Bend	Skip	Wave1	Wave2
Walk	0.95	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Run	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Jump	0.00	0.07	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P-jump	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Jack	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Side	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.06	0.00	0.00
Bend	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Skip	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

In order to quantify the effectiveness of the proposed method, the achieved results are qualitatively compared with those obtained by other investigators. The outcome of the comparison is presented in Table 2. In light of this comparison, one can see that the proposed method is competitive with other state-of-the-art methods [7,38–40]. It might be worth mentioning here that all the methods that we compared our method with, except the method proposed in [7], have used similar experimental setups. Thus, the comparison appears to be meaningful and fair.

Table 2. Comparison with a number of related results reported in recent literature.

Method	Recognition Rate
Our method	98.20%
Fathi and Mori [7]	100.00%
Sadek et al. [41]	97.80%
Bregonzio et al. [39]	96.60%
Zhang et al. [40]	92.80%
Niebles et al. [42]	90.00%

It should finally be noted that all of the experiments reported in this work have been carried out on a 3.2 GHz Intel dual core machine with 4 GB of RAM, running Microsoft Windows 7 Professional. The recognition system has been implemented by using Microsoft Visual Studio 2013 Professional Edition (C++) and OpenCV Library for feature detection and classification. As a final remark, we emphasize that the most significant feature of the proposed approach is its rapidity; our action recognizer runs comfortably in real-time in almost all of the experiments (i.e., at roughly 24 fps on average). This supports the expectation that the proposed method can be used in real-world settings and is amenable to working with real-time applications.

5. Conclusions

In this paper, a discriminative framework for action recognition has been presented, based on a novel feature descriptor so-called 2D f-HOL for action skeleton and a discriminative LDCRF model for feature classification. The proposed approach has been evaluated on the popular Weizmann dataset. The obtained results have demonstrated that the approach has a competitive performance compared to most existing approaches, with an average recognition rate of 98.2%. Moreover, the approach works very efficiently, and it is able to be applicable in real-time scenarios. An important aspect of future work will involve further validation of the approach on more realistic datasets presenting many technical challenges in data handling, such as object occlusion and significant background clutter.

Acknowledgments: This research was supported by the Transregional Collaborative Research Center SFB/TRR 62 "Companion Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). We would also like to thank anonymous reviewers for their thorough reading of the manuscript and their insightful comments and suggestions.

Author Contributions: All authors have contributed substantially to the work reported in this paper and they have read and approve the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Laptev, I.; Pérez, P. Retrieving actions in movies. In Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV), Rio de Janeiro, Brazil, 14–21 October 2007.
2. Sadek, S.; Al-Hamadi, A.; Michaelis, B.; Sayed, U. Human Action Recognition via Affine Moment Invariants. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR'12), Tsukuba, Japan, 11–15 November 2012; pp. 218–221.
3. Sadek, S.; Al-Hamadi, A.; Michaelis, B.; Sayed, U. Human Action Recognition: A Novel Scheme Using Fuzzy Log-Polar Histogram and Temporal Self-Similarity. *EURASIP J. Adv. Signal Process.* **2011**, *2011*, 1–9.
4. Johansson, G. Visual motion perception. *Sci. Am.* **1975**, *232*, 76–88.
5. Sadek, S.; Al-Hamadi, A. Vision-Based Representation and Recognition of Human Activities in Image Sequences. Ph.D. Thesis, Otto-von-Guericke-Universität Magdeburg (IIKT), Magdeburg, Germany, 2013.
6. Aggarwal, J.K.; Cai, Q. Human motion analysis: A review. *Comput. Vis. Image Underst.* **1999**, *73*, 428–440.
7. Fathi, A.; Mori, G. Action recognition by learning mid-level motion features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; IEEE: Piscataway, NJ, USA, 2008.
8. Sadek, S.; Al-Hamadi, A.; Michaelis, B.; Sayed, U. An SVM approach for activity recognition based on chord-length-function shape features. In Proceedings of the IEEE International Conference on Image Processing (ICIP'12), Orlando, FL, USA, 30 September–3 October 2012; pp. 767–770.
9. Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Washington, DC, USA, 17–21 October 2005; Volume 2, pp. 1395–1402.
10. Chakraborty, B.; Bagdanov, A.D.; Gonzàlez, J. Towards Real-Time Human Action Recognition. *Pattern Recognit. Image Anal.* **2009**, *5524*, 425–432.
11. Sadek, S.; Al-Hamadi, A.; Michaelis, B.; Sayed, U. Towards Robust Human Action Retrieval in Video. In Proceedings of the British Machine Vision Conference (BMVC'10), Aberystwyth, UK, 31 August–3 September 2010; pp. 1–11.
12. Sadek, S.; Al-Hamadi, A.; Michaelis, B.; Sayed, U. A Statistical Framework for Real-Time Traffic Accident Recognition. *J. Signal Inf. Process. (JSIP)* **2010**, *1*, 77–81.
13. Efros, A.A.; Berg, A.C.; Mori, G.; Malik, J. Recognizing action at a distance. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 2, pp. 726–733.

14. Sullivan, J.; Carlsson, S. Recognizing and tracking human action. In Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark, 28–31 May 2002; Volume 1, pp. 629–644.
15. Nowozin, S.; Bakir, G.; Tsuda, K. Discriminative subsequence mining for action classification. In Proceedings of the Eleventh IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1919–1923.
16. Liu, J.; Shah, M. Learning human actions via information maximization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008.
17. Little, L.; Boyd, J.E. Recognizing people by their gait: The shape of motion. *Int. J. Comput. Vis.* **1998**, *1*, 1–32.
18. Cutler, R.; Davis, L.S. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. PAMI* **2000**, *22*, 781–796.
19. Shechtman, E.; Irani, M. Space-time behavior based correlation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 405–412.
20. Bobick, A.; Davis, J. The Recognition of Human Movement Using Temporal Templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267.
21. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008.
22. Fernando, B.; Gavves, E.; Oramas, M.J.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5378–5387.
23. Kantorov, V.; Laptev, I. Efficient feature extraction, encoding, and classification for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 2593–2600.
24. Ikizler, N.; Forsyth, D. Searching video for complex activities with finite state models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 17–22 June 2007.
25. Olivera, N.; Garg, A.; Horvitz, E. Layered representations for learning and inferring office activity from multiple sensory channels. *Comput. Vis. Image Underst.* **2004**, *96*, 163–180.
26. Blei, D.M.; Lafferty, J.D. Correlated topic models. *Adv. Neural Inf. Process. Syst.* **2006**, *18*, 147–154.
27. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
28. Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the SIGIR99 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; pp. 50–57.
29. Zivkovic, Z. Improved adaptive Gausian mixture model for background subtraction. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 23–26 August 2004.
30. Bar-Shalom, Y.; Li, X.R. *Estimation and Tracking: Principles, Techniques, and Software*; Artech House: Boston, MA, USA, 1993.
31. Ahmad, M.; Lee, S.W. Human Action Recognition Using Shape and CLG-motion Flow From Multi-view Image Sequences. *J. Pattern Recognit.* **2008**, *7*, 2237–2252.
32. Bakheet, S.; Al-Hamadi, A. A Hybrid Cascade Approach for Human Skin Segmentation. *Br. J. Math. Comput. Sci.* **2016**, *17*, 1–18.
33. Matas, J.; Galambos, C.; Kittler, J.V. Robust Detection of Lines Using the Progressive Probabilistic Hough Transform. *Comput. Vis. Image Underst.* **2000**, *78*, 119–137.
34. Deufemia, V.; Risi, M.; Tortora, G. Sketched symbol recognition using Latent-Dynamic Conditional Random Fields and distance-based clustering. *Pattern Recognit.* **2014**, *47*, 1159–1171.
35. Ramírez, G.A.; Fuentes, O.; Crites, S.L.; Jimenez, M.; Ordonez, J. Color Analysis of Facial Skin: Detection of Emotional State. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 474–479.
36. Morency, L.P.; Quattoni, A.; Darrell, T. Latent-dynamic discriminative models for continuous gesture recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 17–22 June 2007.

37. Lafferty, J.; McCallum, A.; Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
38. Sadek, S.; Al-Hamadi, A.; Michaelis, B.; Sayed, U. An Efficient Method for Real-Time Activity Recognition. In Proceedings of the International Conference on Soft Computing and Pattern Recognition (SoCPaR'10), Cergy Pontoise/Paris, France, 7–10 December 2010; pp. 7–10.
39. Bregonzio, M.; Gong, S.; Xiang, T. Recognising action as clouds of space-time interest points. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–26 June 2009.
40. Zhang, Z.; Hu, Y.; Chan, S.; Chia, L.T. Motion context: A new representation for human action recognition. In Proceedings of the 10th European Conference on Computer Vision (ECCV), Marseille, France, 12–18 October 2008; Volume 4, pp. 817–829.
41. Sadek, S.; Al-Hamadi, A.; Krell, G.; Michaelis, B. Affine-Invariant Feature Extraction for Activity Recognition. *ISRN Mach. Vis. J.* **2013**, *1*, 1–7.
42. Niebles, J.; Wang, H.; Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vis.* **2008**, *79*, 299–318.